DOCUMENT RESUME

ED 479 553 IR 022 159

AUTHOR	Jansen, Bernard J.
TITLE	A Preliminary Mapping of Web Queries Using Existing Image Query Schemes.
PUB DATE	2002-10-00
NOTE	9p.; In: E-Learn 2002 World Conference on E-Learning in Corporate, Government, Healthcare, & Higher Education. Proceedings (7th, Montreal, Quebec, Canada, October 15-19, 2002); see IR 022 093.
AVAILABLE FROM	Association for the Advancement of Computing in Education (AACE), P.O. Box 3728, Norfolk, VA 23514. Tel: 757-623-7588; e-mail: info@aace.org; Web site: http://www.aace.org/DL/.
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	EDRS Price MF01 Plus Postage. PC Not Available from EDRS
DESCRIPTORS	Classification; Computer System Design; Indexing; Information Retrieval; Information Systems; *Online Searching; Search Engines; Search Strategies; World Wide Web
IDENTIFIERS	*Digital Imagery; *Query Formulations; Query Processing

ABSTRACT

End user searching on the Web has become the primary method of locating images for many people. This study investigates the nature of Web image queries by attempting to map them to known image classification schemes. In this study, approximately 100,000 image queries from a major Web search engine were collected in 1997, 1999, and 2001. A stratified sample of the queries was classified using three previously published image classification schemes. The results indicated that the features and attributes of Web image queries differ relative to image queries utilized in other information retrieval systems and user populations. Give the growing ubiquitous nature of the Web, the implications for the design of information retrieval systems and repositories are discussed. (Contains 17 references.) (Author)

Reproductions supplied by EDRS are the best that can be made
from the original document.

A Preliminary Mapping of Web Queries Using Existing Image Query Schemes

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

G.H. Marks

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Bernard J. Jansen
School of Information Sciences and Technology
The Pennsylvania State University
001 Thomas Bldg
University Park, PA, 16801, USA
Phone: 814-865-6459 Fax: 814-865-6426
Email: jjansen@acm.org

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

This document has been reproduced as received from the person or organization originating it.

Minor changes have been made to improve reproduction quality.

Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

ABSTRACT: End user searching on the Web has become the primary method of locating images for many people. This study investigates the nature of web image queries by attempting to map them to known image classification schemes. In this study, approximately 100,000 image queries from a major web search engine were collected in 1997, 1999, and 2001. A stratified sample of the queries was classified using three previously published image classification schemes. The results indicated that the features and attributes of web image queries differ relative to image queries utilized in other IR systems and user populations. Given the growing ubiquitous nature of the web, the implications for the design of information retrieval systems and repositories are discussed.

ED 479 553

Introduction

The tremendous growth in the quantity of digital images is driving the need for more effective methods of storing, searching, and retrieving image data. With the development of such technologies, image indexing and retrieval have received increasing attention. However, there has been little research into image searching on the Web, especially how users search for images in the web environment. Given the pervasiveness of the web, the manner in which user's conduct image searches on web may also be good indications of how they will conduct future searches in other environments.

One can classify current image retrieval approaches, whether on traditional systems or on the web, as either concept or content based (Rasmussen, 1997). In the concept-based approach, image retrieval research has focused on the retrieval of images utilizing indexed image collections (Enser & McGregor, 1992; Goodrum and Spink, 1999). Researchers and practitioners have created thesauri for visual information, such as the Library of Congress Thesaurus for Graphics Materials, Metadata (Greenberg, 1993). These thesauri are then used to index images within a collection.

Conversely, in the content-based approach, the focus has been on indexing images at the pixel level and the implementation of search features using pixel comparison (Wang, 2000). Content-based image IR systems allow users to search image collections using color, texture, shape, and spatial similarity. These systems many times also provide text-based search functions for notations and text descriptions embedded within images. New technologies have placed the emphasis on content-based retrieval, with commercial systems such as the MediaSite.com (http://www.mediasite.com) system.

However, it is not clear how the retrieval functionality of either of these approaches systems correlates with the image needs of real users. Concerning the context approach, Brown, Hidderly, Griffin, and Rollason (1996) point out that experts are not a good source of terms that are preferred by real users. Concerning the content approach, Chen (2001) notes that users seldom search using content characteristics. Although user studies within specific domains have been conducted, Jansen and Pooch (2001) have shown that Web users differ in their interaction with

IR022159

BEST COPY AVAILABLE

IR systems relative to users in other environments. In this respect, studies of the image needs of web users are especially needed for effective and efficient design of web-based, image IR systems.

The research presented in this article focuses on the information needs of web user's as represented by their image queries. Specifically, web user queries from a major web search engine are investigated. We analyze the end-user's image query structure across three periods (1997, 1999, and 2001), classify user queries using accepted classification schemes, compare the characteristics to those identified in previous studies and provides suggestions for improving image indexing and retrieval systems to meet the information need of web users.

Review of Literature

User Information Needs and Image Queries

Ellis (1984, p. 25) advocates that design of IR systems must be with "an orientation to the use and the user's interaction with information sources. Fidel [8] points out that studies in image retrieval have ignored several fundamental issues, including the differences between text and image retrieval, the image attributes important for retrieval, and the characteristics of user's image queries. These points are especially germane with respect to image searching on the web given the almost complete lack of studies in this area.

There have been some studies of image searching on the web, including Goodrumand Spink (1999) and Jansen, Goodrum and Spink (2000) that have analyzed images queries from web users as part of a review of multimedia queries. However, these studies presented primarily factual characteristics of image queries; none of these studies attempted theoretical classification of web image queries using known classification schemes. Without such research, one cannot determine whether or not the web user population's image information needs conforms to currently accepted image retrieval approaches.

Image Retrieval on the Web

In general, Web users must search for multimedia information as they would search for textual information (Schauble, 1997) with many Web IR systems providing no special mechanism for multimedia searching. The advantage of this approach is that image searching is performed in an identical manner to text searching. No additional burden is placed on the searcher. If the searcher desires an image document, the searcher enters the query and specifies some image attribute. The disadvantage of this approach is that it places more contextual knowledge burden on the searchers, who may not be familiar with image formats. Necessitating that users translate a non-textual information need into a textual query may challenges a user's cognitive load, creating a semantic gap (Gudivada & Raghavan, 1995). Excite (http://www.excite.com) and Yahoo search engine (http://www.yahoo.com) are two such Web IR systems that take this approach.

Some Web IR systems specialize in image collections or provide mechanisms for users searching for images (e.g., radio boxes or media specific search syntax). Alta Vista (http://www.altavista.com) and Lycos (http://www.lycos.com) searchers can narrow a query to specifically search for an image. ClipArt Searcher (http://www.webplaces.com/search/), the Yahoo Picture Gallery (http://gallery.yahoo.com), and Webseek (http://www.ctr.columbia.edu/webseek/) are examples of image specific retrieval applications allowing users to search by terms or select from general categories of images and video. Both Webseek and Alta Vista return thumbnail images and file names in the document result list. Webseek also provides tools for content-based searching for images and videos using color histograms generated from the visual scenes.

In terms of image indexing, most web IR systems follow a relative straight forward approach. The simplest image search algorithm locates images files by searching for file extensions, matching the filename to terms in the query (Witten, Moffat, & Bell, 1994). This may not be a good approach, especially given the move to automatic generation of image file names by applications (e.g., converting Power Point © presentations to Web pages or transferring files from a digital camera). In these cases, the file name has little to nothing to do with the image. Web IR systems may also retrieve documents that are primarily textual but that contain embedded images. In these cases, the image filename may not match the query terms, but the Web documents may contain text that does match the query terms. The assumption being that the image relates to the text.

Given the inadequate research of user image needs on the web and a lack of theoretical background for the design and evaluation of web image databases (Chen, 2001), it is logical to examine the both web image queries and present classification schemes to discover the utility of the models for the design of web-based image retrieval systems. A review of three, previously published, image query classification methodologies follow.

Enser and McGregor Classification Categories

Enser and McGregor (1992) analyzed 2,722 image queries requests on an image system with approximately half a million engravings, woodcuts, drawings, cartoons, and maps. The combined size of the collection was about 10 million images. The researchers classified these 2,722 requests in four categories: (1) Unique, (2) Unique with refiners, (3) Non-unique, (4) Non-unique with refiners. The property of uniqueness is readily applied to a request for the visual representation of an entity where the desired occurrence can be differentiated from very other occurrence of the same entity type (Enser and McGregor, 1992, p. 17). Using refiners, one can narrow these queries some degree of specifity. The researchers point out that almost 70% (1905) of the queries were for a unique person, object or event, and 34% (925) of the queries were refined, mostly by time.

Jorgensen 12 Image Attribute Classes

Jorgensen (1998) conducted a study with 107 first semester library science students with a variety of undergraduate majors. The study participants viewed 6 projected color images that were randomly selected from *The Society of Illustrators 25th Annual of American Illustration*. The participants described the 6 projected images, and Jorgensen then classified the descriptions into the 12 classes of image attributes.

Chen's Categories and Attributes Modifications

Chen (2001) collected 534 queries from 29 college students majoring in art history, using pre-search and post-search questionnaires. The art course students utilized an image collection containing over 75,000 volumes and subscriptions to over 200 journals in relevant fields. The collection also contained significant holding in several art related fields and facsimile collections of medieval manuscripts and original scrolls. Chen did not, however, focus on whether or not the queries were utilized on an image retrieval system. Chen mapped the queries to previous query classification methods, including Enser and McGregor (1992) categories, combining some of Jorgensen (1998) attribute classes. Based on Chen's analysis, the researcher proposed modifications to Enser and McGregor (1992) categories.

The next section of the paper describes the research design used in this study.

Research Design

This research reports the analysis of image queries from a major web search engine. The objective was to examine these image queries from real web users with real information needs, mapping these queries to known image query classification features.

Research Questions

The following research question is addressed: Can web image queries be effectively mapped using previously identified image classification schemes?

Background

The Web is an immense repository of image information (Lesk, 1997). The hypertext transfer protocol (HTTP) lends itself to the easy transfer of audio, video, and image formats integrated with textual information. Lawrence and Giles (1999) estimated there are 180 million images on the publicly indexed Web and 3Tb of image data, not including other types of multimedia files, such as audio and video. As such, the web is an immense, heterogeneous image collection.

Research Procedure

A total of approximately 100,000 image queries from the Excite search engine (http://www.excite.com) were collected in 1997, 1999, and 2001. Each transaction record contained three fields. With these three fields, one can locate a user's initial query and recreate the chronological series of actions by each user in a session:

1. **Time of Day**: measured in hours, minutes, and seconds from 00:00:01.
2. **User Identification**: an anonymous user code assigned by the server.
3. **Query Terms**: exactly as entered by the given user.

Overall, the metrics used in this research are as defined in (Jansen & Pooch, 2001). We defined the session, query, and term variables as:

1. **Session**: A session is the entire series of queries by a user over time, from one query to many.
2. **Query**: A query consists of one or more search terms, and may include logical operators and modifiers.
3. **Term**: A term is any unbroken string of characters (i.e. a series of characters with no space between any of the characters). The characters in terms included everything – letters, numbers, and symbols.

Founded in 1994, Excite, Inc. was a major Internet media public company that offers free Web searching and a variety of other services. It has since been purchased and is no longer an independent company. Excite searches were based on the exact terms that a user enters in the query; however, capitalization is disregarded, with the exception of logical commands AND, OR, and AND NOT. Stemming is not available. An online thesaurus and concept linking method called Intelligent Concept Extraction (ICE) was used, to find related terms in addition to terms entered. Search results were provided in a ranked relevance order. A number of advanced search features were available. A page of search results contained ten answers at a time ranked as to relevance. For each site, the results list provided the title, URL (Web site address), and a summary of its contents. A user could click on the title to go to the Web site. A user could also click for the next page of ten answers.

Content Analysis

A set of 587 image queries was randomly selected from one of the transaction logs. The image queries were then categorized using each of the three classifications schemes discussed previously. A query could be mapped to more than one feature or attribute. Therefore, the number of mapping incidents may be greater than the number of queries. We now present the findings of our research.

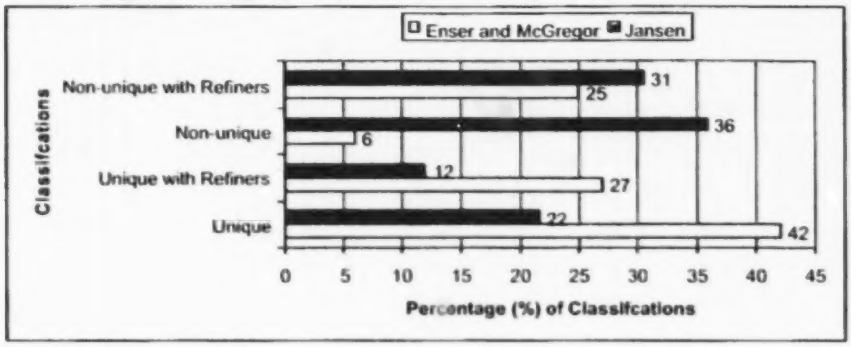
Research Findings

We analyzed to what degree does the image queries map to previous identified image query features, categories, and image attributes. A set of 583 image queries was randomly selected from one of the transaction logs and then mapped to each of following classification schemes.

Mapping Results using Enser and McGregor's Categories

The image queries were mapped to Enser and McGregor's image categories. The overall results are displayed in Figure #1.

Figure 1: Mapping using Enser and McGregor's Four Categories

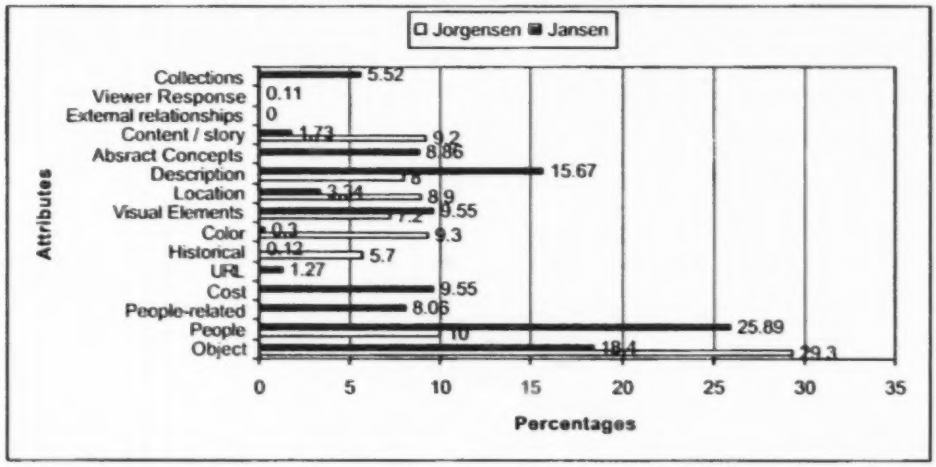


The total number of judgments for the study of Enser and McGregor's was 583. From Figure #1, one can see that the percentage of requests for unique image information was substantial lower on the web than reported by Enser and McGregor's, and naturally, the percentage of non-unique request was higher. Perhaps surprising, given the antidotal comments concerning web users, the use of refiners was quite high, comparable and even exceeding that reported by Enser and McGregor.

Mapping Results using Jorgensen's Image Attributes

We classified the image queries to Jorgensen's image attributes. The overall results are displayed in Figure #2.

Figure 2: Mapping using Jorgensen's image attributes.



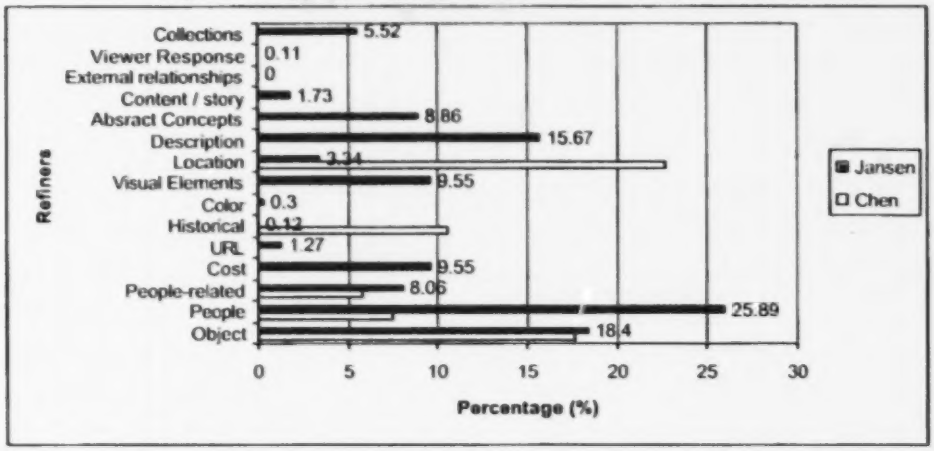
The total number of judgments for the study of Jorgensen's image attributes was 869. Jorgensen's original 12 attributes did not effectively describe the attributes of the web image queries, so an additional three attributes were added, namely cost, URL, and collections. Concerning the three attributes added, it is apparent that cost is factor with web users, a fact that would not be apparent in lab studies using students, as Jorgensen and others have done. Given the unique hypertext environment, the location of the images is apparently a concern of Web users given the high number of URL requests and searches for images collections. People and people – related queries seem to dominate the information needs of web users. Content-based attributes, such as color, do not appear to be utilized by web users even though web IR systems provide relevant results to these content-based attributes (e.g., using Alta Vista, search for blue images).

BEST COPY AVAILABLE

Mapping based on Chen's Modifications

We classified the image queries to Chen image categories modifications. The overall results are displayed in Figure #3.

Figure 3: Mapping using Chen image categories modifications



The total number of judgments for the study of Chen's image attributes modifications was 869. Similar to Jorgensen's scheme, we had to add attributes to account for differences in the web query. The results from the web mapping was much more diverse than that reported by Chen.

Discussion

In Enser and McGregor's study (1992), approximately 70% of the queries were for a unique person object or event. Also, 34% of these requests were refined, primarily by time. In applying the Enser and McGregor's categories to web image queries, we find that the majority of queries were non-unique, but almost all of these were refined in some manner, with cost (24%), presentation aspects (20%), or pornographic (16%) refiners being the most common.

In the Jorgensen study (1998), the literal Object class held the highest position in the searching task (29.3%), followed by the people class (10.0%), and then the color class (9.3%). In the analysis of web queries, the most frequency occurring attribute is People (26%), followed by Description (16%) and then Object (18%).
In the Chen study [4], the 29 searchers mainly utilized unique terms and seldom used refiners. Chen identified five refiners, which were Object, People, Historical, and Location. The web queries correspond to these five categories but there are also substantial numbers of cost (10%), Visual Elements (10%), Description (16%), Abstract Concepts (9%), and Collections (6%).

Conclusion

How users search for images on web is certain to have an impact on how users interaction with other knowledge bases. Thanks in part to the Web, end user searching is now the primary method of locating images for many people. This study investigated web image queries. By mapping a sub-set of image queries to known image classification schemes, the results indicated that the features and attributes of web image queries differ relative to image queries utilized in other IR systems and user populations. The majority of image queries were requests for non-unique entities and almost all contained some refiner. Cost is the most prevalent refiners. Most image queries are for people. Despite these differences, it is apparent that known image classification schemes can be modified to account for the characteristics of web users.

BEST COPY AVAILABLE

In future research, larger sub- sets of web images queries will be provided to multiple reviewers in order to obtain inter rater reliabilities. In this respect, we model off of the technique used by (Chen, 2001). We expect this technique to refine our classification attributes. From this, one can derive a meaningful scheme for classifying web image queries.

References

Brown, P., Hiderley, R., Griffin, H., & Rollason, S. (1996). The democratic indexing of image. The New Review of Hypermedia and Multimedia. 2, 107 – 121.

Chen, H. (2001). An Analysis of Image Queries of Art History. Journal of the American Society for Information Science and Technology, 52(3), 260 – 273.

Ellis, D. (1984). Theory and Explanation in Information Retrieval Research. Journal of Information Science, 8, 25 – 38.

Enser, P.G.B. & McGregor, C.G. (1992). Analysis of Visual Information Retrieval Queries. Report on Project G16412 to the British Library Research and Development Department. London: British Library.

Goodrum, A. & Spink, A. (1999). Visual Information Seeking: A study of image queries on the World Wide Web. *Proceedings the 1999 Annual Meeting of the American Society for Information Science*, Washington, DC. November, 1999.

Greenberg, J. (1993). Intellectual Control of visual Archives: A Comparison between the Art and Architecture Thesaurus and the Library of Congress Thesaurus for Graphic and Materials. Cataloging and Classifications Quarterly, 16(1), 85 – 101.

Gudivada, V.V., & Raghavan, V.V. (1995). "Content-based image retrieval systems," *IEEE Computer*, 28(9), 18-22.

Jansen, B. J. and Pooch, U. (2001). Web user studies: A review and framework for future work. *Journal of the American Society of Information Science and Technology*. 52(3), 235 - 246.

Jansen, B. J., Goodrum, A., and Spink. A. (2000). Searching for multimedia: video, audio, and image Web queries. *World Wide Web Journal*, 3(4), 249 - 254.

Jorgensen, C. (1998). Attributes of images in describing tasks. Information Processing and Management, 34(2/3), 161 – 174.

Lawrence, S., & Giles, C. L. (1999). Accessibility of information on the web. *Nature*. 400, 107 –109.

Lesk, M. (1997). Going digital. *Scientific American, 276*(3), 58-60.

Rasmussen, E.M. (1997). Indexing Images. Annual Review of Information Science and Technology, 32, 169 – 196.

Saracevic, T. (1997). Users lost: Reflections on the past, future, and limits of information science. *SIGIR Forum, 31*(2) 16-27.

Schauble, P. (1997). *Multimedia information retrieval*. Kluwer: Boston.

Wang, J. Z. (2000). "SIMPLIcity: A region-based image retrieval system for picture libraries and biomedical image databases," *Proceedings of. ACM Multimedia*, pp. 483-484, Los Angeles, ACM, October 2000.

Witten, I.H., Moffat, A., & Bell, T. C. (1994). *Managing gigabytes: Compressing and indexing documents and images*. Van Nosstrand Reinhold: New York.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

[X] This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

[ ] This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").